Data in Brief 32 (2020) 106298

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Synthetic database of space objects encounter events subject to epistemic uncertainty

Luis Sánchez*, Massimiliano Vasile

University of Strathclyde, United Kingdom

ARTICLE INFO

Article history:
Received 5 August 2020
Revised 31 August 2020
Accepted 4 September 2020
Available online 9 September 2020

Keywords:
Space traffic management
Evidence theory
Risk assessment
Epistemic uncertainty
Collision risk assessment

ABSTRACT

The databases included on this article refers to variables and parameters belonging to the Space Traffic Management (STM), Evidence Theory and Machine Learning (ML) fields. They have been used for implementing ML for autonomously predict risk associated to a close encounter between two space (Sanchez and Vasile, *On the Use of Machine Learning and Evidence Theory to Improve Collision Risk Management,* Acta Astronautica, Special Issue for ICSSA2020, In Press [1]). The position of the objected is assumed to be affected by epistemic uncertainty, which has been modeled according to Dempster-Shafer Evidence theory (DSt) [2].

Six datasets are presented. Two (*DB1* and *DB2*, respectively) include samples of space object close encounters subject to epistemic uncertainty on the relative position. Other two databases (*DB3* and *DB4*, respectively) include the values of the Cumulative Plausibility and Belief Curves (*CPC* and *CBC*, respectively) of each sample included in *DB1*. The remaining databases (*DB5* and *DB6*), contain the value of the *CPC* and *CBC* of each sample included in *DB2*. All of them are synthetic databases created using computer simulation to obtain the results presented in [1]. *DB1* database is constituted by 9,000 samples and 45 columns and a header, while *DB2* is formed by 28,800 samples and 45 columns and a header. These databases come from a set of, respectively, 5 and 14 different families of encounter geometries defined by the range of values that can be assigned to the bounds of the intervals for the uncertain variables, assumed to be

DOI of original article: 10.1016/j.actaastro.2020.08.004
* Corresponding author.
E-mail address: luis.sanchez-fdez-mellado@strath.ac.uk (L. Sánchez).

https://doi.org/10.1016/j.dib.2020.106298

affected by epistemic uncertainty, considered to have been provided by two sources of information. The uncertain variables are: the miss distance, $[\mu_x, \mu_y]$, on the impact plane (B plane), the standard deviation of the relative position projected on the B plane, $[\sigma_x, \sigma_y]$, and the Hard Body Radius of the combined objects, *HBR*. The dataset is completed with STM related parameters: miss distance and covariance matrix of the uncertain ellipse projected on the B plane enclosing all samples defined by the uncertainty intervals, the Probability of Collision (P_C) of this ellipse or the elapsed time to the Time of Closest Approach (*TCA*); with DSt related parameters: Belief and Plausibility of certain values of *Pc*; and the class of the event according to the classification detailed in [1]. *DB3* and *DB4* are constituted by 34 columns and 9000 rows containing the Plausibility and Belief for *Pc* values and the corresponding Probabilities of Collision necessary to build the *CPC* and CBC of the events in *DB1*, while *DB5* and *DB6* are constituted by 34 columns and 28,800 rows containing the Plausibility and Belief for *Pc* values and the corresponding Probabilities of Collision values necessary to build the *CPC* and CBC of the events in *DB2*.

These databases have a potential usage by the ML community interested in STM as well as for the space community, especially, space operators interested in introduce epistemic uncertainty on collision risk assessment. These databases contribute to build a scarce field such as the databases of encounter events [3].

Specifications Table

Subject	Aerospace Engineering
Specific subject area	Space Traffic Management and Space Collision Risk Assessment.
Type of data	Table
How data were acquired	Database obtained by software simulation using SW code developed by the authors developed in open source Python code.
Data format	Raw
Parameters for data collection	Data were obtained through computer simulation. A Windows Operative System was used. The code included (a whole version of the code is available in a GitHub repository [4]), developed by the authors, allowed the creation of individual samples and the computation of relevant parameters included on the Databases. All the code was developed in open source Python language and run using Anaconda Prompt.
Description of data collection	Data were collected from computer simulation. A set of 22 parameters were established as initial condition for each sample. These parameters were the bounds of the intervals provided by hypothetical sources of information from certain variables affected by epistemic uncertainty. These variables were randomly generated for each sample in such a way that the bounds of the intervals belong to certain limits associated to encounter geometry families. Parameters related with space risk assessment and with Dempster–Shafer theory of evidence were computed using the aforementioned Python code.
Data source location	Institution: University of Strathclyde City: Glasgow Country: United Kingdom
Data accessibility	Repository name: Synthetic Database of Space Objects Encounter Events Subject to Epistemic Uncertainty. (Mendeley Data) Data identification number: 10.17632/gxk2c45xb7.1 Direct URL to data: http://dx.doi.org/10.17632/gxk2c45xb7.1
Related research article	L. Sánchez, M. Vasile, *On the Use of Machine Learning and Evidence Theory to Improve Collision Risk*, Acta Astronautica, Special Issue for ICSSA2020. In Press.

Value of the Data

- First, data are useful for replicating the results presented in [1]. Second, they are useful in a wider extend. As indicated in [3], the application of ML techniques on STM is limited, in part, due to the lack of adequate databases to train the models. These databases contribute to filling this gap.
- The data can benefit, in the first place, the researchers or entities interested on replicate the study developed in [1]. In the second place, the Machine Learning community, specially, the growing one interested on the Space sector. Finally, the Space Engineering community, specifically, the Space Traffic Management area and those interested on account for epistemic uncertainty on the risk assessment.
- These data can be used as they are presented to replicate work developed on [1] or go further on that study. They can be combined with further STM related data to generate a wider database for further development in this area of study. They can be used by the ML community for test models both supervised (regression and classification) and not supervised (i.e. clustering).
- The inclusion of epistemic uncertainty on collision risk assessment is currently very limited. These databases can contribute to set a benchmark to include criteria and standards that account for this effect using Dempster–Shafer theory of evidence.

1. Data Description

Six different Databases are presented: *Database_System1.csv, Database_System2.csv, Plausibility_System1.txt, Belief_System1.txt, Plausibility_System2.txt*, and *Belief_System2.txt*, named, respectively, *DB1, DB2, DB3, DB4, DB5* and *DB6*. The two first databases share the same configuration, differing in the number of elements and how they have been obtained. *DB3* and *DB5* share also the same structure, differing in the number of rows, similarly to *DB4* and DB6. All the databases are explained below.

1.1. Database_System1.csv (DB1) *and* Database_System2.csv *(*DB2*)*

These databases include information about space risk collision assessment under probabilistic and epistemic uncertainties [1]. They contain parameters related to Space Collision Risk Assessment, like miss distance, relative position standard deviation and covariance matrix, time to Time of Closest Approach (*TCA*) and Probability of Collision (P_C), and also include Evidence Theory related parameters, such as Plausibility (*Pl*) and Belief (*Bel*) for values of P_C, limits of intervals of the uncertain variables and basic probabilistic assignment (*bpa*) values. The two databases differ on the set of encounter geometries from where bounds of the intervals of the uncertain variables were withdrawn as initial conditions to compute the rest of the parameters (more details on how these parameters are obtained are given in the next Section).

DB1 is comprised of 9000 rows and 45 columns. Database 2 is comprised of 28,800 rows and 45 columns. Each row on both Databases represents a space encounter event. All the units are in International System unless otherwise stated. The columns on both Databases represent:

- Column 1: ID of the encounter event: *ID_event*.
- Columns 2-11: upper and lower bounds of the intervals for each of the uncertain variables (standard deviation on both axis of B plane $[\sigma_x, \sigma_y]$, Hard Body Radius (*HBR*) and miss distance components on the B plane $[\mu_x, \mu_y]$) provided by the source of information Source 1: $\underline{\sigma_{1x}}$, $\overline{\sigma_{1x}}$, $\underline{\sigma_{1y}}$, $\overline{\sigma_{1y}}$, $\underline{HBR_1}$, $\overline{HBR_1}$, $\underline{\mu_{1x}}$, $\overline{\mu_{1x}}$, $\underline{\mu_{1y}}$, $\overline{\mu_{1y}}$.
- Column 12: Basic Probabilistic assignment of Source 1: bpa_1.
- Column 13–22: upper and lower bounds of the intervals for each uncertain variables provided by the source of information Source 2: $\underline{\sigma_{2x}}$, $\overline{\sigma_{2x}}$, $\underline{\sigma_{2y}}$, $\overline{\sigma_{2y}}$, $\underline{HBR_2}$, $\overline{HBR_2}$, $\underline{\mu_{2x}}$, $\overline{\mu_{2x}}$, $\underline{\mu_{2y}}$, $\overline{\mu_{2y}}$.
- Column 23: Basic Probabilistic assignment of Source 2: bpa_2.

- Column 24–25: Components in the B plane reference frame of the center of the uncertain ellipse including all the set of ellipses that belong to the intervals specified in Columns 2–11 and Columns 13–22, weighted with the values of Columns 12 and 23: μ_{0x}, μ_{0y}.
- Columns 26–29: Covariance matrix of the ellipse indicated in the previous bullet point, expressed in the B plane reference frame: σ_{0xx}, σ_{0xy}, σ_{0yx}, σ_{0yy}.
- Columns 30–31: Components of the center of the uncertain ellipse indicated in the two previous bullet points expressed in a reference frame aligned with the principal axis of the ellipse and center in the same point of the B plane reference frame: $\hat{\mu}_{0x}$, $\hat{\mu}_{0y}$.
- Columns 32–33: Covariance matrix (diagonal) of the previous uncertain ellipse expressed in a reference frame aligned with the principal axis of the ellipse: $\hat{\sigma}_{0xx}, \hat{\sigma}_{0yy}$.
- Columns 34: Probability of collision computed using Eq. (1), of the uncertain ellipse associated with Source 1, which includes all the set of ellipses defined by the intervals included in Columns 2–11: P_{C1}.
- Columns 35: Probability of collision computed using Eq. (1), of the uncertain ellipse associated with Source 2, which includes all the set of ellipses defined by the intervals included in Columns 13–22: P_{C2}.
- Columns 36: Probability of collision computed using Eq. (1), of the uncertain ellipse specified in Columns 26–33: $\widehat{P_C}$.
- Columns 37–38: Plausibility and Belief of the value of Probability of Collision indicated in Column 36 according to the values included in Columns 2–23, compute as in [1]: $Pl(\widehat{P_C})$, $Bel(\widehat{P_C})$.
- Columns 39 and 40: Plausibility and Belief at a specific threshold for Probability of Collision ($P_{C0} = 4.4 \times 10^{-4}$) according to the values included in Columns 2–23, compute as in [1]: $(P_{c0}), Bel(P_{C0})$.
- Columns 41–42: Values of Probability of Collision where the Cumulative Plausibility and Belief Curves, respectively, computed based on the values in Column 2–23, jump above (or below) the respective thresholds for Plausibility ($Pl_0 = 0.5$) and Belief ($Bel_0 = 0.5$): $P_C(Pl_0), P_c(Bel_0)$.
- Column 43: Flag indicating if there is any value of Probability of Collision in the interval $[\widehat{P_C}, 1]$ at which the difference between Belief and Plausibility is greater than a threshold ($\Delta = 0.5$). *0* means gap is always below, *1* means the gap is greater than the threshold at some value of P_c.
- Column 44: Elapsed time to the Time of Closest Approach (*TCA*) in days.
- Column 45: Class indicating the risk of the conjunction event accounting for Epistemic Uncertainty and time to the *TCA*, according to *Evidence-based Classification Criterion 3* in [1].

1.2. Plausibility_System1.txt (DB3) *and* Plausibility_System2.txt *(*DB5*)*

These databases include the values of Probability of Collision and Plausibility needed to build the Cumulative Plausibility Curves (*CPC*). *DB3* includes the *CPC* of the events in *DB1* and it is constituted by 9000 rows. *DB5* includes the *CPC* of the events in *DB2* and it is constituted by 28,800 rows. Each sample on both databases is constituted by three rows:

- First row of the geometry: ID of the geometry. Note that each geometry on *DB3* and Plausibiltiy_*System2.txt* is associated with three events in *DB1* and *DB2*, respectively. Thus, *event_ID* = 0, 1 and 2 in *DB1* are associated to the same *geometry_ID* = 0 in *DB3*.
- Second row of the geometry: values of P_C where Plausibility changes value. Note *CPC* is a step function, thus these P_C values indicate the limits of the horizontal segment.
- Third row of the geometry: values of the Pl at each P_C of the previous row.

1.3. Belief_System1.txt *(*DB4*) and* Belief_System2.txt (DB6)

These databases include the values of Probability of Collision and Belief needed to build the Cumulative Belief Curves (*CBC*). The structure is like *DB3* and *DB5*, but instead of Plausibility, it

includes Belief:

- First row of the geometry: ID of the geometry, which coincides with ID on the Plausibility databases. Note that each geometry on *DB4* and *DB6* provides three events in *DB1* and *DB2*, respectively. Thus, *event_ID* = 0, 1 and 2 in *DB1* are associated to the same *geometry_ID* = 0 in *DB4*.
- Second row of the geometry: values of P_C where Belief changes value. Note *CBC* is a step function, thus these P_C values indicate the limits of the horizontal segment.
- Third row of the geometry: values of the Bel at each P_C of the previous row.

1.4. Main.py *and* CreateDB.py

The code files *Main.py* and *CreateDB.py* include in the same repository as the databases were used to generate these them. *Main.py* is the main file where the parameters of the simulation are set and the calls to the other functions are made. *CreateDB.py* includes the different functions to compute the parameters included in the databases.

2. Experimental Design, Materials and Methods

Data are obtained via computer simulation. All the databases share the starting point: the 2 sets of intervals, defined by the lower and upper bounds, provided by two hypothetical different sources of information (*Source 1* and *Source2*) of the five uncertain variables considered: σ_x, σ_y, *HBR*, μ_x, μ_y. More details on the interpretation of these variables and sources of information can be found in [1]. These intervals defined the encounter geometry between two space objects, expressed in the impact plane or B plane, where the knowledge on the position of the objects is subject to epistemic uncertainty. This uncertainty is captured with the intervals accordingly to Dempster-Shafer theory of evidence [2]. The values of the bounds of the intervals are limited to certain ranges. Each of these ranges of values has been defined to create a certain geometry family [1]. In addition to the intervals, a Basic Probabilistic Assignment (*bpa*) is assigned to each source of information regarding its reliability. To create *DB1*, and thus *DB3* and *DB4*, 5 different families of geometries, detailed in Table 3 in [1], have been defined and a total of 9,000 specific encounter geometries have been defined, from where three individual encounter events have been created by assigning them three different times to *TCA*. To create *DB2*, and hence *DB5* and *DB6*, the same samples included in *DB1* have been taken in addition to other 1,800 samples per each of the 5 families mentioned before and another 9,000 samples from 9 new geometry families (1,200 samples per family) detailed in Table 5 in [1], to make a total of 28,800 samples. The interval bounds of the uncertain variables and the *bpa* of both sources are included in *DB1* and *DB2*. The function *create_intervals* in *CreateDB.py* performs this task.

The intervals of the uncertain variables represent a set of uncertain ellipses projected on the B plane. From this set of ellipses, a bigger one enclosing all of the samples included on the set of ellipses defined by the intervals can be defined, characterized by the distance to the center of the B plane (miss distance) and its covariance matrix. These values, expressed in the B plane reference frame and in a reference frame aligned with the principal axis of the ellipse, are included in *DB1* and *DB2*. The Probability of Collision (P_c) is a metric used in Collision Risk Assessment [5]. It can be understood as the area below the distribution defined by the uncertain ellipse overlapping the Hard Body Radius (*HBR*) of the combined objects. In [5], an expression for its calculation is provided, Eq. (1). The Probability of Collision of the ellipse enclosing all set of ellipses provided by Source 1 (P_{C1}), the Probability of Collision of the ellipse enclosing all the samples defined by the set of ellipses indicated by Source 2 (P_{C2}) and the Probability of Collision of the ellipse enclosing the previous set of ellipses ($\hat{P}_C$) are included

in *DB1* and *DB2*. The function *FittedEllipse* in *CreateDB.py* computed the ellipse and the values of P_c.

$$P_C = \frac{1}{2\pi\sigma_x\sigma_y}\iint_{HBR} exp\left[-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{\left(y-\mu_y\right)^2}{\sigma_y^2}\right)\right]dxdy \tag{1}$$

The Probability of Collision is a probabilistic metric assuming uncertain variables as purely aleatory. For accounting for the epistemic uncertainty expressed in the form of the aforementioned intervals, the Plausibility and Belief concepts from Evidence Theory [2] can be obtained to create the Cumulative Plausibility Curve (*CPC*) and Cumulative Belief Curve (*CBC*) as explained in [1]. These curves present the Plausibility and Belief versus the Probability of Collision and indicate, respectively, the lack of support against a value of P_C and the positive support to a value of P_C according to the evidence (intervals and *bpa*). The values of Plausibility and the values of Pc at which Plausibility jumps are saved in *DB3* and *DB5*. The values of Belief and the values of Pc at which Belief jumps are saved in *DB4* and *DB6*. The values of the Plausibility and Belief at $\hat{P}_C$ and at certain threshold of Probability of Collision, P_{C0}, are included on *DB1* and *DB2*, as well as the values of Probability of Collision, $P_C(Pl)$ and $P_C(Bel)$, at which Plausibility and Belief jumps above (or below) the corresponding threshold, Pl_0 and Bel_0, respectively. A detailed explanation of the calculation of the *CPC* and *CBC* is included in [1]. The function *PlBel_curves* in *CreateDB.py* computed the *CPC* and *CBC*, while the function *PlBel_values* obtained the key values of these curves.

A flag is included in *DB1* and *DB2* indicating if the difference between Plausibility and Belief at any value of Probability of Collision greater than $\widehat{P_C}$ is larger than a certain value, Δ. In case there is any value where the difference is bigger, the flag takes the value *1*, otherwise takes the value *0*. Each of the encounter geometries defined by the uncertain intervals can be associated with a time to the Time of Closest Approach, indicating the proximity of the encounter event. Three times have been associated with each of the geometries: one less than 2 days, another bigger than 4 days, and a third one between 2 and 4 days. This time is contained in *DB1* and *DB2*. Finally, the event defined by the uncertain intervals provided by the Sources and the *bpa* assigned to each of them, along with the time to the *TCA* and the Plausibility and Belief of the key values of Probability of Collision allows a classification of the risk the encounter event present that takes into account the epistemic uncertainty, or lack of information, on the position of the bodies as explained in [1] in the so-called Evidence-based Criterion 3, detailed in Table 1. The function *CompleteDB* in *CreateDB.py* performed these tasks as well as to put together all the parameters in a .csv file.

Note that some function included on the GitHub[1] repository mentioned above may be needed to the computation of certain parameters included on these databases.

Table 1
Evidence-bases event classification Criterion 3.

Time to TCA	$P_C(Bel_0)$	*Degree of confidence at Pc0 Pl(Pc0) – Bel(Pc0)*	*Class*
$t_{TCA} < T_1$	$P_C(Bel_0) \geq P_{C0}$	–	*1*
	$P_C(Bel_0) < P_{C0}$	$Pl(P_{C0}) - Bel(P_{C0}) \leq \Delta$	*5*
		$Pl(P_{C0}) - Bel(P_{C0}) > \Delta$	*1*
$T_1 \leq t_{TCA} < T_2$	$P_C(Bel_0) \geq P_{C0}$	–	*2*
	$P_C(Bel_0) < P_{C0}$	$Pl(P_{C0}) - Bel(P_{C0}) \leq \Delta$	*5*
		$Pl(P_{C0}) - Bel(P_{C0}) > \Delta$	*3*
$t_{TCA} \geq T_2$	$P_C(Bel_0) \geq P_{C0}$	–	*2*
	$P_C(Bel_0) < P_{C0}$	$Pl(P_{C0}) - Bel(P_{C0}) \leq \Delta$	*4*
		$Pl(P_{C0}) - Bel(P_{C0}) > \Delta$	*3*

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

The authors would like to express their appreciation to the Royal Aeronautical Society for having funded their attendance to the 2nd International Conference on Space Situational Awareness, ICSSA-2020, where this data were first presented along, and the University of Strathclyde for financing the research behind the generation of the databases.

References

[1] L. Sanchez, M. Vasile, On the use of machine learning and evidence theory to improve collision risk, Acta Astronaut., Special Issue for ICSSA2020. (In Press).
[2] G. Shafer, A Mathematical Theory of Evidence, 1st ed., Princeton University Press, Princeton, New Jersey, 1976.
[3] L. Sanchez L., M. Vasile, E. Minisci, AI and space safety: collision risk assessment. In: Schrogl KU. (ed.) *Handbook of Space Security*. Springer, Cham.
[4] GitHub repository SMART (Strathclyde Mechanical and Aerospace Research Toolboxes: https://github.com/strath-ace
[5] S. Alfano, Review of conjunction probability methods for short-term encounters, Adv. Astronaut. Sci. 127 (2007) 719–746.